Réduction itérative du biais pour des lisseurs multivariés

Pierre-André Cornillon, Nick Hengartner, Eric Matzner-Løber

Réduction itérative du biais pour des lisseurs multivariés

Pierre-André Cornillon[a], Nicolas Hengartner[b] & Eric Matzner-Løber[c]

[a] *: UMR MISTEA - Montpellier SupAgro, 34060 Montpellier Cedex 1, France ;*
[b] *: Los Alamos National Laboratory, NW, USA ;*
[c] *: Univ. Rennes, 35043 Rennes, France.*

Mots-clés : Modèles semi et non paramétriques, Statistique mathématique

Résumé

La méthode IBR (iterated biased reduction) permet d'estimer une fonction de régression m inconnue lorsque les variables explicatives sont à valeurs dans $\mathbb{R}^d$. Pour estimer la fonction m, les méthodes non-paramétriques classiques souffrent du fléau de la dimension. En pratique, il faut donc supposer des hypothèses structurelles : modèles additifs, modèles à directions révélatrices... A contrario IBR estime directement la fonction de régression m. Elle concurrence MARS, les directions révélatrices ou les modèles additifs et sur des exemples réels ou simulés et elle apporte des gains significatifs sur l'erreur de prévision. Cette méthode utilise en pratique un lisseur pilote soit de type splines plaque-minces soit de type noyau gaussien. Cet estimateur pilote est utilisé de manière répétée afin d'estimer le biais et permet de l'enlever progressivement. La méthode, à l'instar du L_2 boosting, nécessite donc l'estimation de l'itération optimale. Des résultats de vitesse de convergence (vitesse minimax) de l'erreur quadratique moyenne de l'estimateur (avec itération optimale) ont été obtenus. L'optimalité du critère de choix de l'itération (GCV) a aussi été démontré. Un exemple simulé simple ($d = 2$) et un exemple réel ($d = 8$) seront traités et comparés aux méthodes existantes : GAM, MARS, PPR, ou L_2-boosting. Un package R disponible sur le CRAN permet d'utiliser cette méthode très simplement.

Abstract

This paper presents a general procedure for nonparametric multivariate regression smoothers that outperforms existing procedures such as MARS, additive models, projection pursuit or L_2 additive boosting on both real and simulated datasets. In multivariate nonparametric analysis, sparseness of the covariates also called curse of dimensionality, forces one to use large smoothing parameters. This leads to biased smoothers. We still propose to use classical nonparametric linear smoothers, such as thin plate splines or kernel smoothers, but instead of focusing on optimally selecting the smoothing parameter, we fix it to some reasonably large value to ensure an over-smoothing of the data. The resulting (base) smoother has a small variance but a substantial bias. Afterward, we propose to iteratively correct the biased initial estimator by an estimate of the bias

obtained by smoothing the residuals. In univariate settings, we relate our procedure to L_2-Boosting. Rules for selecting the optimal number of iterations are also proposed and, based on empirical evidence, we propose one stopping rule. In the regression framework, when the unknown regression function m belongs to the Sobolev space $\mathcal{H}^{(\nu)}$ of order ν, we show that using a thin plate splines base smoother and the proposed stopping rule leads to an estimate $\hat{m}$ which converges to the unknown function m. Moreover, our procedure is adaptive with respect to the unknown order ν and converges at the minimax rate. We apply our method to both simulated and real data and show that our method compares favourably with existing procedures such as MARS, additive models, L_2 boosting or projection pursuit, with improvement on means squared error up to 30%. An R package is available on CRAN.

1 Introduction

Si nous souhaitons expliquer une variable Y par un ensemble de d variables explicatives $X_1, \cdots, X_d$, la régression constitue un outil classique de la statistique. Cette famille de modélisation comprend la régression paramétrique linéaire ou non-linéaire, la régression non-paramétrique utilisant des lisseurs construits à partir d'ondelettes, de noyaux, de splines [voir par exemple 7] etc.

Dès que le nombre d'observations est modéré (de l'ordre de plusieurs centaines) et que le nombre de variables d est plus grand que 3 ou 4, les approches non-paramétriques classiques rencontrent le problème dit du fléau de la dimension. Dans ce cas, un modèle structurel est souvent utilisé : par exemple un modèle additif [4], des directions révélatrices ou MARS [5].

Le boosting est aussi une réponse possible au problème de régression [voir 3] et cette méthode possède maintenant de nombreux développements comme adaboost, logitboost pour la discrimination ou le L_2 boosting pour la régression. Cette dernière peut être utilisée avec de nombreux lisseurs et donne lieu à une modélisation additive par composante [voir 1].

Un lien entre les méthodes de L_2 boosting et la régression non-paramétrique existe via la réduction itérative de biais. Cette idée trouve ses sources dans la méthode du twicing de [8] où un estimateur pilote est corrigé via une estimation de son biais. L'idée d'itérer cette méthode est aussi évoquée dans la discussion de [6] sur l'interprétation statistique du boosting.

Dans ce résumé nous rappelons brièvement le concept de réduction itérative du biais et nous rappelons son lien avec le L_2 boosting en univarié ($d = 1$). Nous rappelons ensuite les propriétés théoriques obtenues en multivarié et nous concluons sur des exemples.

2 Réduction itérative du biais

Soit le modèle de régression suivant

$$Y_i \;=\; m(X_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

où $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $m(\cdot)$ est une fonction lisse inconnue et les erreurs ε_i sont des variables aléatoires indépendantes des covariables, indépendantes entre elles, de moyenne nulle et de variance constante σ^2. Il est plus aisé de réécrire l'équation (1) sous forme vectorielle en posant $Y = (Y_1, \ldots, Y_n)^t$, $m = (m(X_1), \ldots, m(X_n))^t$ et $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^t$, donnant ainsi

$$Y \;=\; m + \varepsilon. \tag{2}$$

Les lisseurs linéaires permettent d'obtenir une estimation de m via

$$\hat{m}_1 = S_\lambda Y, \tag{3}$$

où S_λ est une matrice de lissage $n \times n$ (paramétrée par λ) et où $\hat{m}_1 = \hat{Y} = (\hat{Y}_1, \ldots, \hat{Y}_n)^t$ représente le vecteur des ajustements. Ce type de lisseur inclue les splines de régression, les smoothing splines, les splines plaque-minces, les noyaux de Nadaraya-Watson ou les polynômes locaux. Nous allons nous intéresser principalement aux noyaux (λ est alors la fenêtre) et aux splines plaque-minces (où λ représente la pénalité).

Le lisseur linéaire (3) a pour biais et variance

$$B(\hat{m}_1) = E[\hat{m}_1|X] - m = (S_\lambda - I)m \qquad V(\hat{m}_1|X) = \sigma^2 S_1 S_1'.$$

Pour estimer le biais il suffit de voir que les résidus $R_1 = Y - \hat{m}_1 = (I - S_\lambda)Y$ ont pour espérance $E[R_1|X] = m - E[\hat{m}_1|X] = (I - S_\lambda)m = -B(\hat{m}_1)$, ce qui suggère d'estimer le biais par un lissage de l'opposé des résidus

$$\hat{b}_1 := -S_\lambda R_1 = -S_\lambda (I - S_\lambda) Y.$$

Rappelons qu'en multivarié ($d > 1$), avec un nombre modéré d'observations, les lisseurs de type noyaux ou splines plaque-minces se heurtent au fléau de la dimension : comme le nombre d'observations au voisinage d'un point est très faible il est nécessaire d'utiliser une grande valeur de λ. Ceci mène mécaniquement à un estimateur pilote S_λ biaisé. En conséquence la correction de biais apparaît comme un outil naturel et nécessaire en multivarié. Bien évidemment, si λ est grand tout le biais n'est pas supprimé dès la première correction et une idée naturelle consiste donc à itérer ces étapes de réduction de biais. Ainsi après $k-1$ étapes de réduction de biais nous obtenons le lisseur linéaire à l'itération k :

$$\begin{aligned} \hat{m}_k &= S_\lambda Y + S_\lambda (I - S_\lambda) Y + \cdots + S_\lambda (I - S_\lambda)^{k-1} Y \\ &= (I - (I - S_\lambda)^k) Y. \end{aligned} \tag{4}$$

Quand d vaut un, en univarié, nous obtenons simplement le L_2 boosting (sans shrinkage). Par contre, en multivarié l'équivalence ne tient plus car le L_2 boosting implémente des modèle additifs par composante ou component-wise additive models [voir 2] et ne produit donc pas une estimation directe de m mais une estimation par modèle contraint.

2.1 Propriétés théoriques

Soit Ω un ouvert de $\mathbb{R}^d$ et supposons que la fonction inconnue m de régression appartient à l'espace de Sobolev d'ordre ν noté $\mathcal{H}^{(\nu)}(\Omega)$, où ν est un entier tel que $\nu > d/2$. Soit S la matrice de lissage obtenue avec des splines plaque-minces d'ordre $\nu_0 \leq \nu$ (en pratique nous prenons $\nu_0 = \lfloor d/2 \rfloor + 1$) et $\lambda_0 > 0$ fixé à une valeur raisonnablement élevée.

Un premier théorème montre qu'il existe un nombre d'itérations $k = k(n)$, dépendant de la taille de l'échantillon, pour lequel l'estimateur $\hat{m}_k$ atteint la vitesse minimax. Ce résultat généralise celui obtenu par [2] en univarié. Cependant, afin qu'il soit utile il est nécessaire de proposer une règle d'arrêt et de prouver son utilité. Si nous utilisons le critère GCV, alors nous pouvons montrer que sur les ensembles $\mathcal{K}_n = \{1, \ldots, n^{1+\gamma}\}$, $\gamma \geq 0$, si les erreurs ont un moment d'ordre $4q$ fini (avec $q > (1+\gamma)(2\nu/d+1)$), quand n tend vers l'infini,

$$\frac{\|\hat{m}_{\hat{k}_{GCV}} - m\|^2}{\inf_{k \in \mathcal{K}_n} \|\hat{m}_k - m\|^2} \longrightarrow 1, \quad \text{en probabilité.}$$

Nous avons donc un critère, le critère GCV, qui permet de sélectionner une itération optimale pour les estimateurs pilotes utilisant des splines plaque-minces.

3 Exemples

3.1 Exemple bivarié

Nous allons utiliser un exemple simulé avec la fonction test de Wendelberg [9]

$$\begin{aligned} m(x_1, x_2) &= \frac{3}{4}\exp\left\{-\frac{(9x-2)^2+(9y-2)^2}{4}\right\} + \frac{3}{4}\exp\left\{-\frac{(9x+1)^2}{49} - \frac{(9y+1)^2}{10}\right\} \\ &\quad + \frac{1}{2}\exp\left\{-\frac{(9x-7)^2+(9y-3)^2}{4}\right\} - \frac{1}{5}\exp\left\{-(9x-4)^2-(9y-7)^2\right\}. \quad (5) \end{aligned}$$

qui est représentée en figure 1.

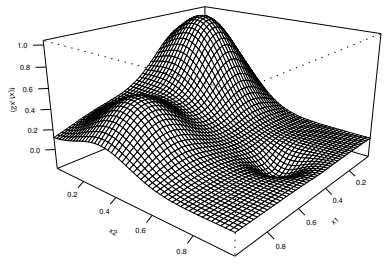

FIG. 1 – Fonction $m(x_1, x_2)$ (5) sur le pavé $[0,1] \times [0,1]$.

Nous utilisons comme échantillon d'apprentissage $n = 100$ observations régulièrement réparties sur la grille $\{0.05, 0.15, \ldots, 0.85, 0.95\}^2$. Les erreurs sont gaussiennes de moyenne nulle et de variance telle que le rapport signal sur bruit vaut 5.

Une séquence de correction itérative de biais pour un lisseur pilote de type splines plaque-minces converge vers l'interpolation (voir figure 2 (c)). Après un nombre adéquat d'itérations, le lisseur obtenu donne un bon estimateur de la fonction (figure 2 (b)).

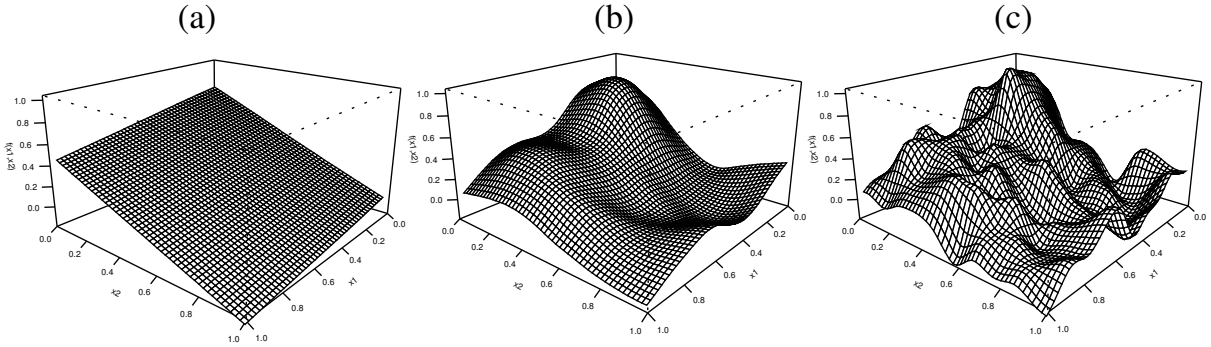

FIG. 2 – Estimateur par réduction itérative de biais obtenu avec 100 observations simulées avec la fonction (5) (voir figure 1) et évalué sur une grille régulière de $[0, 1] \times [0, 1]$. (a) estimateur pilote utilisant des splines plaque-minces (b) estimateur obtenu après 100 itérations (c) estimateur obtenu après 50000 itérations.

Pour comparer notre procédure de réduction itérative de biais (avec choix du nombre d'itérations k par GCV) à la méthode classique des splines plaque-minces (où le paramètre λ est estimé par validation croisée), nous prenons un échantillon test qui est la grille de 50×50 points régulièrement espacés dans $(0, 1)^2$. Sur cet échantillon test nous obtenons une erreur absolue moyenne de 0.0578 pour la réduction itérative de biais alors que l'estimateur classique donne une erreur absolue moyenne de 0.0582. Même sur cet exemple jouet, la procédure se comporte bien.

3.2 Exemple réel

Reprenons l'exemple classique de la concentration en ozone dans le bassin de Los Angeles [2]. Le nombre d'observation est $n = 330$ et le nombre de variables explicatives $d = 8$.

Si nous souhaitons utiliser les splines plaque-minces, l'ordre ν_0 doit être plus grand que $d/2$, c'est-à-dire $\nu_0 = 5$. Le degré de liberté minimum de S_λ est alors $M_0 = 495$ qui est plus grand que n. Les splines plaque-minces ne sont pas utilisables pour cet exemple. Rappelons que la méthode nécessite un lisseur de départ S_λ très lisse, c'est-à-dire avec un degré de liberté faible comparé à n. Même si n était de l'ordre de 500, les splines plaque-minces ne seraient pas satisfaisantes.

Utilisons alors un lisseur de type noyau gaussien obtenu par simple produit de noyaux univariés. Afin que chaque variable explicative soient traitée avec la même importance, la

fenêtre de chaque variable est choisie de sorte que le noyau univarié gaussien atteigne un degré de liberté de 1.1. Ce lissage est implémenté avec le package `ibr` disponible sur le CRAN (par exemple `http://cran.univ-lyon1.fr/`) :

```
> data(ozone)
> res.ibr <- ibr(ozone[,-1],ozone[,1],df=1.1)
```

Si nous séparons le jeu de données en 50 partitions aléatoires de 33 observations en test et 297 en apprentissage, l'erreur quadratique moyenne (MSE) que l'on obtient est de 14.98, erreur que l'on peut comparer aux méthodes classiques comme GAM (package `mgcv` : 17.44), MARS (package `mda` : 17.49), les directions révélatrices (`ppr` : 17.79 avec `nterms=2`) ou le L_2 boosting (package `mboost` : 17.23). Sur cet exemple, la méthode apporte un gain d'environ 15%.

Bibliographie

[1] Bühlmann, P. et Hothorn, T. (2007) Boosting algorithms : regularization, prediction and model fitting (with discussion). *Statistical Science*, 22, 477–505.

[2] Bühlmann, P. et Yu, B. (2003) Boosting with the l_2 loss : Regression and classification. *J. Amer. Statist. Assoc.*, 98, 324–339.

[3] Friedman, J. (2001) Greedy function approximation : A gradient boosting machine. *Ann. Statist.*, 28.

[4] Hastie, T. J. et Tibshirani, R. J. (1990) *Generalized Additive Models*. Chapman & Hall, London.

[5] Hastie, T. J., Tibshirani, R. J. et Friedman, J. H. (2001) *The elements of statistical learning : data mining, inference and prediction*. Springer, New-York.

[6] Ridgeway, G. (2000) Additive logistic regression : a statistical view of boosting : Discussion. *Ann. of Statist.*, 28, 393–400.

[7] Simonoff, J. S. (1996) *Smoothing Methods in Statistics*. Springer, New York.

[8] Tukey, J. W. (1977) *Explanatory Data Analysis*. Addison-Wesley.

[9] Wendelberger, J. (1982) Smoothing noisy data with multivariate splines and generalized cross-validation. *Ph.D thesis, University of Wisconsin*.